Fitting Regression Models To Finite Mixtures

John Williams

Senior Teaching Fellow
Department of Marketing
Otago University

Abstract

Marketing researchers often need to fit regression models to data that may not be ho-mogeneous with respect to the variables involved. Recent developments in finite mixture modeling have provided techniques to partition the data into homogeneous subgroups and perform regression analyses within each group in a simultaneous manner. This paper presents an overview of these developments and simple examples, using both simulated and empirical data, that show the utility of the approach.

The most practical benefit for marketing researchers is that there is evidence that the approach demonstrated here is superior in terms of parameter recovery to the more usual practice of first clustering the data using a method such as K-means and then fitting regression models to each cluster.

Introduction

Often one has data at hand that may be considered as being drawn from a mixture of several homogeneous populations. Indeed, some of the most basic classical statistical tests (for example Student's t test (Student 1908) and Fisher's F test (Fisher 1925)) explicitly test the assumption that the data are drawn from a single homogeneous population. If those tests are rejected one must analyse each population separately with respect to the variables involved.

A simple example is where a market researcher has several semantic differential scales that measure the attractiveness of a human model in a print advertisement. One may reasonably assume that male and female respondents will have differing response distributions, presumably with different location parameters and perhaps different dispersion parameters also.

For example, suppose the scale mentioned above is a seven point scale anchored by "very attractive" and "very unattractive" with the value 4 representing ambivalence. If the responses are Gaussian distributed with equal variance, and male respondents have a mean of 3 on the item while females have a mean of 5, then the overall mean will be 4, which is certainly misrepresentative of the market from a business point of view. This misrepresentation would still occur if one group had a mean response of 1, while the other has a mean response of 7, which could lead to an even worse mistake for a marketing practitioner.

Finite mixture modeling is a tool that can be used (but is not limited to) attack clustering problems and latent class analysis of the nature described above. (Latent class analysis is a subset of finite mixture modeling.) However finite mixture modeling, and in particular regression modeling for finite mixtures, is not a general clustering tool. To use regression models for finite mixtures one usually starts with the regression model and then, in the face of evidence of heterogeneity, one can apply regression in a finite mixture context. The research into identifying finite mixtures of Normal distributions dates back to Karl Pearson's series of seminal papers around the turn of the century (Pearson 1894). More recently several authors (see Titterington, Smith, and Makov (1985) and McLachlan and Bas-ford (1988) for reviews) have examined the problem, particularly as an alternative to cluster analysis. The treatment given in that body of literature concentrates on mixtures on univariateunconditional (i.e. no model imposed on the mean or covariance structure) mixtures and does not address the problem of analysing mean or covariance structures..

Regression models for finite mixtures

Regression models for imposing structures on finite mixtures, known as clusterwise regression models, a term due to Späth (1979), have been proposed by DeSarbo and Cron (1988) for univariate dependent variables. (Further work has since been undertaken to extend this model — see Wedel and DeSarbo (1994) and Wedel and Kamakura (2000) for reviews — but this paper will not deal with these extensions.) DeSarbo and Cron (1988) write their model as:

(1) $$f(y_i|x_i, k) = \sum_{k=1}^{K} \pi_k \theta(Y_i; \gamma_k + \mathbf{II}_k x_i, \sigma_k^2)$$

where π_k k is the mixing proportion of group k, γ_kk is a vector of regression constants (intercepts) for the kth group, $\mathbf{II}_k$ is a 1 x q matrix of regression coefficients (where q is the number of x variables) in the kth group and σ_k is the error variance in the kth group. The function θ is the likelihood function of the data given its mean and error variance, the construction of which will be discussed in below. The subscript $i = 1, \ldots, n$ indicates the ith of n observations.

The model under consideration is that the data arise as independent and identically distributed realisations from a mixture (finite sum) of Gaussian distributions, that is:

(2) $$Y_i = \sum_{k=1}^{K} \pi_k f_{ik}(Y_i|X_{ij}, \sigma_k^2, b_{jk}), j = 1, \ldots, q$$

where

(3) $$f_{ik}(Y_i|X_{ij}, \sigma_k^2, b_{jk}) = (2\pi\sigma_k^2)^{-1/2} \exp\left[\frac{-(Y_i - x_i b_i)^2}{2\sigma_k^2}\right]$$

From this definition of the data generating mechanism the log-likelihood function can be written as:

(4) $$\log L = \sum_{i=1}^{n} \log\left\{\pi_k (2\pi\sigma_k^2)^{-1/2} \exp\left[\frac{-(Y_i - x_i b_k)^2}{2\sigma_k^2}\right]\right\}$$

Given the number of groups and the observed data one may then estimate the parameters of the (log) likelihood function needed for it to be maximised. To do so, it is necessary to impose the following restrictions:

(5) $$0 \leq \pi_k \leq 1$$

(6) $$\sum_{k=1}^{K} \pi_k = 1$$

(7) $$\sigma_k^2 > 0$$

This specification of the model has several important properties. Most importantly, for finite mixtures of Gaussian distributions (unlike mixtures of other distributions) the parameters of the density functions are identified. Secondly, there are no sufficient estimators for the parameters of a Gaussian mixture. Thirdly, consistent estimators do not exist unless the last restriction in (7), that the error variance is strictly positive, is imposed.

The estimates obtained allow one to assign cases to groups using Bayes' rule:

(8) [bloody horrible equation with carets goes here]

and then subject i belongs to group k iff $P_{jk} > P_{jk}$ $\forall l \neq k = l \ldots K$.

As mentioned above the parameters of the density functions of mixture of Gaussian distributions are identified. This depends on a fixed number of groups and does not, unfortunately, imply that a test for the number of groups is identified.

DeSarbo and Cron (1988) use AIC (Akaike 1974) to choose the number of groups, increasing the number of groups and comparing AIC values for each condition. Other approaches are possible, in particular an ad hoc test

proposed by Wolfe (1971), or bootstrapping the likelihood ratio (McLachlan 1987). One cannot use the usual likelihood ratio X^2 test because its regularity conditions are not satisfied. In particular, the possible values of the parameters lie on the boundary of the parameter space, i.e. it is possible that a given mixing proportion is 0 or 1. AIC is based on the likelihood ratio and thus its assumptions are violated also. For this reason tests for the number of groups are (currently) not available, and the usual procedures are employ AIC and the test of Wolfe (1971) as guides only, rather than absolute criteria.

Parameters are estimated by DeSarbo and Cron (1988) using the EM algorithm (Dempster, Laird, and Rubin 1977) and treating information regarding group membership as missing data. In order to perform the computations we need a likelihood function that reflects the restrictions in equation 7. Such a function can be written as:

(9)
$$\phi = \sum_{i=1}^{n} \log \left[\sum_{k=1}^{K} \pi_k f_{ik}(Y_i | X_{ij}, \sigma_k^2, b_{jk}) \right] - \mu \left(\sum \pi_k - 1 \right)_{k=1}^{K}$$

DeSarbo and Cron (1988) show that the estimating equations for maximising the likelihood with respect to the parameters b_{jk} and σ_k^2 are identical to weighted least squares regression where the weights are $\hat{}_{pik}^{1/2}$. Thus in the E stage of the EM algorithm one estimates π_k and, P_{ik}, while in the M stage one estimates b_{ik} and σ_k^2 by weighted least squares regressions.

The iterative nature of the EM algorithm balances the regression and clustering phases, optimising the results from both. Contrast this to the more usual practice of first clustering the data, perhaps using K means, and then performing a regression analysis. Such a procedure stakes the whole outcome of the analysis on the goodness of the clustering algorithm, and ignores information from other groups.

Examples of the procedure

Two examples of the procedure are given below, one using synthetic data and the other using real data. The examples used here are for illustrative purposes only and should not be taken as an attempt to show the superiority of De Sarbo and Cron's method in all situations. In particular it must be emphasised that the method is useful only when the groups are well separated. As the differences in parameters of groups decrease, estimates of parameters and associated standard errors become inconsistent. Day (1969) suggests using the generalised distance (Mahalanobis 1936) and Jedidi et al. (1997) define an entropy measure for the difference in parameters. Yung (1994) found that estimates became inconsistent when Day's measure was below 3 (in a confirmatory factor analysis context).

The first example uses simulated data and an implementation of the algorithm described above, using the statistical programme R (Ihaka and Gentleman 1996) for data generation and MECOSA (Arminger et al. 1996) parameter estimation.

K means clustering has been chosen to contrast clusterwise regression against sequential clustering followed by separate regressions. One could argue that another clustering method might be superior to the EM algorithm used here, but as there are so many clustering methods it would take much more space than this article allows to compare them all. Also, K means is one of the most widely used clustering methods.

In this example 500 observations were simulated for one dependent variable and two independent variables in each of two groups, giving 1000 observations. The independent variables and the error term were all standard Gaussian variables, and the regression parameters areas given in Table 1, where the first subscript denotes the group to which the parameter be-longs. The parameter β_{i1} is the intercept in each group. The Known column gives the results where group membership is known and a separate regression was carried out on each group, whereas EM and K Means give results for when the respective procedures estimated group membership.

Table 1: Parameter estimates for simulated data

Parameter	True	Known	EM	K Means
β_{11}	1.5	1.504	1.507	1.979
β_{12}	0.5	0.528	0.452	0.029
β_{13}	0.8	0.804	0.860	0.310
β_{21}	0.3	0.316	0.394	-0.078
β_{22}	0.2	0.224	0.275	-0.037
β_{23}	-0.4	-0.324	-0.299	-0.157

Both the EM algorithm and K means were run by specifying two groups known *a priori*. In practice this would not be the whole analysis, as one would usually try several different values for the number of groups. However, as mentioned above, the problem of testing for the number of groups has not been completely resolved.

The results show clearly that the EM approach is superior to K means for these data. It must be emphasised once more however that this is merely an example and is not meant to provide a rigorous demonstration of the superiority of De Sarbo and Cron's method in all cases.

The next example uses real data collected from a survey conducted in Germany by a commercial research firm regarding attitudes relating to snack bars. Data relating to a particular brand ("Brand A") are available from 467 respondents. Several items relate to respondents' feelings toward various brands. The items used for this analysis were:

Brand A . . .
...is a brand that I trust more than other brands
...is a brand that I identify with more than with other brands
...is a brand that differentiates itself positively with respect to other brands
...is a brand that I like more than other brands
...is a brand that I intend to buy in the future

The last item measures intention to buy. It is of interest to investigate the relationship

between intention to buy and the attitude items. Demographic variables "Sex" and "Income" were also included in the analysis.

A brief overview of the results is presented below. The results for both an aggregate analysis and a two group solution are shown in Table 2, which gives the estimated regression parameters with standard errors in parentheses. Space considerations preclude a detailed substantive analysis of the data, but a brief synopsis of one possible interpretation is given below.

Interpretation of the aggregate level estimates in Table 2 would lead a market researcher or brand manager to suppose that there is a low general level of intention to buy, but that intention to buy is increased by general liking of the brand; by the ability of the brand to differentiate itself positively from other brands; and by the degree to which consumers are able to relate to the brand. Of these factors, ability to relate to the brand is the most important in terms of intention to buy. Trust in the brand is not important, neither is the sex or income of consumers.

Table 2: Parameter estimates for empirical data

Parameter	Aggregate		Group One		Group Two	
Intercept	0.62	(0.20)	-0.21	(0.26)	1.65	(0.38)
Sex	0.08	(0.09)	-0.03	(0.11)	0.07	(0.14)
Income	-0.03	(0.02)	-0.05	(0.02)	-0.04	(0.02)
Trust	0.10	(0.05)	-0.01	(0.06)	0.24	(0.07)
Identify	0.39	(0.05)	0.65	(0.06)	-0.10	(0.10)
Differentiate	0.22	(0.05)	0.32	(0.07)	0.26	(0.08)
Like	0.15	(0.05)	0.15	(0.07)	0.27	(0.07)
n	467		172		295	

The clusterwise regression coefficients tell a different story. In group one the estimates are similar to the aggregate level estimates, but now it appears there may be a very small sex and income effect, and the intercept is lower. In addition, the magnitude of the effect on the ability of the consumer to identify with the brand is much greater than at the aggregate level, as is the importance of the brand's ability to differentiate itself positively.

In the second, larger, group the intercept is much higher, indicating that this group is more likely to purchase the brand. In contrast to the aggregate level analysis and the Group One estimates, in Group Two the ability of the consumer to identify with the brand has no effect on their intention to buy it in the future. But now the effect of trust in the brand on intention to buy is large, whereas it is estimated to be low or zero at the aggregate level and in Group One. Also the importance of liking the brand is twice as large as in Group One.

These results suggest that the market for this brand may be segmented into at least two groups. In the larger segment intention to buy is based largely on the ability of the brand to inspire trust but in the smaller segment the determining factor is the ability of the brand to create an image that consumers can identify with. These factors are mutually exclusive: trust is not important in Group One and identification is not important in Group Two.

Conclusion

It can be seen from the above examples that the benefits of the finite mixture approach to fitting regression models are considerable if heterogeneity is present in the data. In particular, effects that are present in one group and not another may be averaged over by aggregate level analysis such that their presence is not revealed. Given that most market research is conducted with segmented markets, but the basis for segmentation may be unknown, this is an important consideration for researchers and brand managers.

Some important caveats must be considered. The parameter estimates from the EM algorithm described in this paper are critically dependent on the separation between the groups (in terms of their parameters). Furthermore, tests for the number of groups are not available. Lastly, the theory is based on the assumption of multivariate Gaussian observations, and violations of this assumption will produce biased and inconsistent estimates. For these reasons care must be taken when applying this algorithm.

References

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control AC-19*(6), 716–723.

Arminger, G., J. Wittenberg, and A. Schepers (1996). *MECOSA 3 User Guide*. Friedrichsdor-f/Ts.: ADDITIVE GmbH.

Day, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika 56*(3), 463–474.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society (Series B) 39*, 1–38.

DeSarbo, W. S. and W. L. Cron (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification 5*, 249–282.

Fisher, R. A. (1925). *Statistical Method for Research Workers* (First ed.). Edinburgh: Oliver and Boyd.

Ihaka, R. and R. Gentleman (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics 5*(3), 299–314.

Jedidi, K., H. S. Jagpal, and W. S. DeSarbo (1997). Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Marketing Science 16*(1), 39–59.

Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Science India 12*, 49–55.

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics 36*, 318–324.

McLachlan, G. J. and K. E. Basford (1988). *Mixture Models*. New York: Marcel Dekker.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London A 185*, 71–110.

Späth, H. (1979). Algorithm 39: Clusterwise linear regression. *Computing 22*, 367–373.

Student (1908). The probable error of a mean. *Biometrika* 6(1), 1–25.

Titterington, D. M., A. F. M. Smith, and U. E. Makov (1985). *Statistical Analysis of Finite Mixture Distributions*. Chichester: Wiley.

Wedel, M. and W. S. DeSarbo (1994). A review of recent developments in latent class regression models. In R. P. Bagozzi (Ed.), *Advanced Methods in Marketing Research*, Chapter 10, pp. 352–388. Cambridge, Massachusetts: Blackwell.

Wedel, M. and W. A. Kamakura (2000). *Market Segmentation: Conceptual and Methodological Foundations* (2nd ed.). International Series in Quantitative Marketing. Boston: Kluwer Academic Publishers.

Wolfe, J. H. (1971). A Monte Carlo study of the sampling distribution of the likelihood ratio for mixtures of multinormal distributions. Technical Bulletin STB 72-2, Naval Personnel and Training Research Laboratory, San Diego, CA.

Yung, Y.-F. (1994). *Finite mixtures in Confirmatory Factor-Analytic Models*. Ph. D. thesis, Department of Psychology, University of California, Los Angeles.